Impact of Quality of Bayesian Network Parameters on Accuracy of Medical Diagnostic Systems

Agnieszka Onisko[1,3] and Marek J. Druzdzel[1,2]

[1] Faculty of Computer Science, Białystok University of Technology, Wiejska 45A, 15-351 Białystok, Poland
[2] Decision Systems Laboratory, School of Information Sciences and Intelligent Systems Programs, University of Pittsburgh, Pittsburgh, PA 15260, USA
[3] Magee-Womens Hospital, Department of Pathology, University of Pittsburgh Medical Center, Pittsburgh, PA 15213, USA

Abstract. While most knowledge engineers believe that the quality of results obtained by means of Bayesian networks is not too sensitive to imprecision in probabilities, this remains a conjecture with only modest empirical support. We summarize the results of several previously presented experiments involving Hepar II model, in which we manipulated the quality of the model's numerical parameters and checked the impact of these manipulations on the model's accuracy. The chief contribution of this paper are results of replicating our experiments on several medical diagnostic models derived from data sets available at the Irvine Machine Learning Repository. We show that the results of our experiments are qualitatively identical to those obtained earlier with Hepar II.

1 Introduction

Decision-analytic methods provide a coherent framework for modeling and solving decision problems in decision support systems [12]. A valuable modeling tool for complex uncertain domains, such as those encountered in medical applications, is a Bayesian network [19], an acyclic directed graph quantified by numerical parameters and modeling the structure of a domain and the joint probability distribution over its variables. There exist algorithms for reasoning in Bayesian networks that compute the posterior probability distribution over some variables of interest given a set of observations. As these algorithms are mathematically correct, the ultimate quality of their results depends directly on the quality of the underlying models and their parameters. These parameters are rarely precise, as they are often based on subjective estimates or data that do not reflect precisely the target population.

The question of sensitivity of Bayesian networks to precision of their parameters is of much interest to builders of intelligent systems. If precision does not matter, rough estimates or even qualitative "order of magnitude" estimates that are typically obtained in the early phases of model building, should be sufficient

without the need for their painstaking refinement. Conversely, if network results are sensitive to the precise values of probabilities, a lot of effort has to be devoted to obtaining precise estimates.

There is a popular belief, supported by anecdotal evidence, that Bayesian network models are tolerant to imprecision in their numerical parameters. Pradhan *et al.* [20] were the first to describe an experiment in which they studied the behavior of a large medical diagnostic model, the CPCS network [15, 23]. Their key experiment, which we will subsequently refer to as the *noise propagation experiment*, focused on systematic introduction of noise in the original parameters (assumed to be the gold standard) and measuring the influence of the amount of noise on the average posterior probability of the true diagnosis. They observed that this average was insensitive to even very large amounts of noise. The noise propagation experiment, while ingenious and thought provoking, offers room for improvements. The first problem, pointed out by Coupé and van der Gaag [7], is that the experiment focused on the average posterior rather than individual posterior in each diagnostic case and how it varies with noise, which is of most interest. The second weakness is that the posterior of the correct diagnosis is by itself not a sufficient measure of model robustness. Practical model performance will depend on how these posteriors are used. In order to make a rational diagnostic decision, for example, one needs to know at least the probabilities of rival hypotheses (and typically the joint probability distribution over all disorders). Only this allows for weighting the utility of correct against the dis-utility of incorrect diagnosis. If the focus of reasoning is differential diagnosis, it is of importance to observe how the posterior in question compares to the posteriors of competing disorders. Another problem is that noise introduced in parameters was assumed to be random, which may not be a reasonable assumption. It is known, for example, that human experts often tend to be overconfident [16]. Yet another opportunity for improvement is looking at precision of parameters rather than their random deviations from the true value. Effectively, the results of the noise propagation experiment are tentative and the question whether actual performance of Bayesian network models is robust to imprecision in their numerical parameters remains open.

Search for those parameters whose values are critical for the overall quality of decisions is known as sensitivity analysis. Sensitivity analysis studies how much a model output changes as various model parameters vary through the range of their plausible values. It allows to get insight into the nature of the problem and its formalization, helps in refining the model so that it is simple and elegant (containing only those factors that matter), and checks the need for precision in refining the numbers [16]. Several researchers proposed efficient algorithms for performing sensitivity analysis in Bayesian networks (e.g., [3, 6, 7, 14]). It is theoretically possible that small variations in a numerical parameter cause large variations in the posterior probability of interest. Van der Gaag and Renooij [11] found that practical networks may indeed contain such parameters. Because practical networks are often constructed with only rough estimates of probabilities, a question of practical importance is whether overall imprecision

in network parameters is important. If not, the effort that goes into polishing network parameters might not be justified, unless it focuses on their small subset that is shown to be critical.

In this paper, we report the results of a series of experiments in which we manipulate the quality of parameters of several real or realistic Bayesian network models and study the impact of this manipulation on the precision of their results. In addition to looking at symmetric noise, like in the original noise propagation experiment, we enter noise in the parameters in such a way that the resulting distributions become biased toward extreme probabilities, hence, modeling expert overconfidence in probability estimates. Our results show that the diagnostic accuracy of Bayesian network models is sensitive to imprecision in probabilities. It appears, however, that it is less sensitive to overconfidence in probabilities than it is to symmetric noise. We also test the sensitivity of models to underconfidence in parameters and show that underconfidence in parameters leads to more error than symmetric noise.

We examine also a related question: "Are Bayesian networks sensitive to precision of their parameters?" Rather than entering noise into the parameters, we change their precision, starting with the original values and rounding them systematically to progressively rougher scales. This models a varying degree of precision of the parameters. Our results show that the diagnostic accuracy of Bayesian networks is sensitive to imprecision in probabilities, if these are plainly rounded. However, the main source of this sensitivity appears to be in rounding small probabilities to zero. When zeros introduced by rounding are replaced by very small non-zero values, imprecision resulting from rounding has minimal impact on diagnostic performance.

Our experiments suggest that Bayesian networks may be less sensitive to the quality of their numerical parameters than previously believed. While noise in numerical parameters starts taking its toll almost from the very beginning, there is a noticeable region of tolerance to small amounts of noise.

The remainder of this paper is structured as follows. Section 2 introduces the models used in our experiments. Section 3 describes our experiments based on introducing noise into probabilities. Section 4 describes our experiments based on progressive rounding of parameters. Finally, Section 5 summarizes our results and main insights obtained from these results.

2 Models studied

The main model used in our experiments is the Hepar II model [18]. This is one of the largest practical medical Bayesian network models available to the community, carefully developed in collaboration with medical experts and parametrized using clinical data.[4] We would like to note that the results for the Hepar II network presented in this paper have been presented before [9, 10, 17]. In addition, we selected three data sets from the Irvine Machine Learning Repository:

[4] Readers interested in Hepar II can download it from Decision Systems Laboratory's model repository at `http://genie.sis.pitt.edu/`.

Table 1. Medical data used in our experiments

data set	instances	variables	variable types	classes
Acute Inflammation	120	8	categorial, integer	4
SPECT Heart	267	22	categorial	2
Cardiotocography	2,126	23	categorical, real	3
HEPAR II	699	70	categorial, real	11

(1) Acute inflammation [8], (2) SPECT Heart [4], and (3) Cardiotocography [22]. Table 1 presents basic characteristics of the selected data sets, including Hepar data. Table 2 presents basic statistics of Bayesian network models that we created from the data. All models consist of only discrete nodes with all continuous variables discretized before the models were learned.

Table 2. Bayesian network models used in our experiments

model	nodes	arcs	states	parameters	avg in-degree	avg outcomes
Acute Inflammation	8	15	17	97	1.88	2.13
SPECT Heart	23	52	46	290	2.26	2.00
Cardiotocography	22	63	64	13,347	2.86	2.91
HEPAR II	70	121	162	2,139	1.73	2.24

Similarly to Pradhan *et al.* [20], for the purpose of our experiments, we assumed that the model parameters were perfectly accurate and, effectively, the diagnostic performance achieved was the best possible. Of course, in reality, the parameters of the model may not be accurate and the performance of the model can be improved upon. In our experiments, we study how this baseline performance degrades under the condition of noise and inaccuracy.

We define diagnostic accuracy as the percentage of correct diagnoses on real patient cases. This is obviously a simplification, as one might want to know the sensititivity and specificity data for each of the disorder or look at the global quality of the model in terms of AUC (Area Under the Curve) of the ROC (Receiver Operating Characteristics) curve, as suggested by a reviewer. This, however, is complicated in case of models focusing on multiple disorders — there is no single measure of performance but rather a measure of performance for every single disorder. We decided thus to focus on the percentage of correct diagnoses.

Because Bayesian network models operate only on probabilities, we assume that each model indicates as correct the diagnosis that is most likely given evidence. When testing the accuracy of models, we were interested in both (1) whether the most probable diagnosis indicated by the model is indeed the correct diagnosis, and (2) whether the set of w most probable diagnoses contains the correct diagnosis for small values of w (we chose a "window" of w=1, 2, 3,

and 4). The latter focus is of interest in diagnostic settings, where a decision support system only suggest possible diagnoses to a physician. The physician, who is the ultimate decision maker, may want to see several top alternative diagnoses before focusing on one.

3 Noise in parameters

Our first series of experiments focused on sensitivity of accuracy of Bayesian network models to symmetric noise in their parameters. When introducing noise into model parameters, we used the approach proposed by Pradhan *et al.* [20], which is transforming each original probability into log-odds form, adding symmetric Gaussian noise parametrized by a parameter σ, and transforming it back to probability, i.e.,

$$p' = Lo^{-1}[Lo(p) + \text{Normal}(0, \sigma)] \ , \tag{1}$$

where

$$Lo(p) = \log_{10}[p/(1-p)] \ . \tag{2}$$

This guarantees that the transformed probability lies within the interval $(0, 1)$.

3.1 Symmetric noise

In [17], we performed experiments focusing on how symmetric noise (see the top two graphs in Figure 2 to get an idea of what this noise amounts to) introduced into network parameters affects the diagnostic accuracy of HEPAR II. Figure 1 presents the diagnostic accuracy of 30 versions of the network (each for a different standard deviation of the noise $\sigma \in < 0.0, 3.0 >$ with 0.1 increments) on the set of test cases for different values of window size as a function of σ.

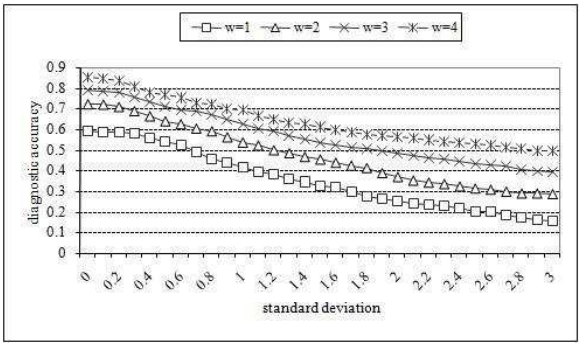


Fig. 1. The diagnostic accuracy of the model under symmetric noise as a function of σ (w=1) [17].

Diagnostic performance seems to deteriorate for even smallest values of noise, although it has to be said that the plot shows a small region (for σ smaller than roughly 0.2) in which performance loss is minimal.

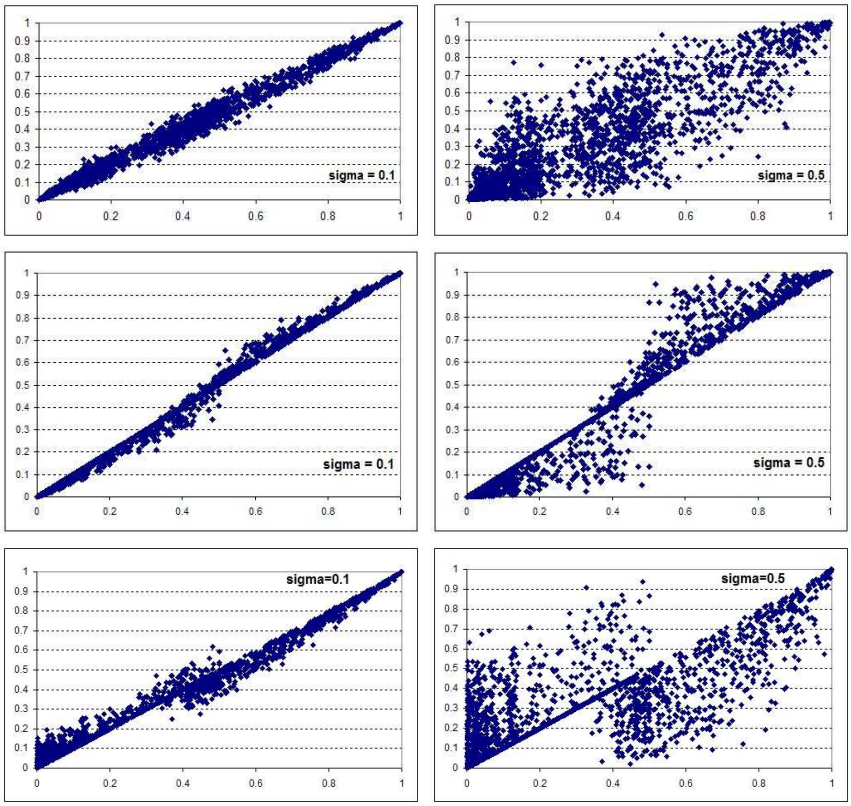


Fig. 2. Scatterplots of the original (horizontal axis) vs. transformed (vertical axis) probabilities for $\sigma = 0.1$ and $\sigma = 0.5$. The top two plots show symmetric noise, the middle two plots show overconfidence, the bottom two plots show underconfidence.

3.2 Biased noise

Symmetric random noise does not seem to be very realistic. It is a known tendency of experts to be overconfident about their probability estimates, i.e., offer more extreme probability estimates than warranted by objective evidence [13, 16]. One way of simulating bias in expert judgment is to distort the original parameters so that they become more extreme (this amounts to modeling expert overconfidence) or more centered, i.e., biased towards uniform probabilities

(this amounts to modeling expert underconfidence). Our next experiment (reported in [9]) focused on investigating the influence of biased noise in HEPAR II's probabilities on its diagnostic performance.

We introduced bias into noise in the following way. Given a discrete probability distribution Pr, for overconfidence, we identified the smallest probability p_S. We transformed this smallest probability p_S into p'_S by making it even smaller, according to the following formula:

$$p'_S = Lo^{-1}[Lo(p_S) - |\text{Normal}(0, \sigma)|] \ .$$

We made the largest probability in the probability distribution Pr, p_L, larger by precisely the amount by which we decreased p_S, i.e.,

$$p'_L = p_L + p_S - p'_S \ .$$

An alternative way of introducing biased noise suggested to us is by means of building a logistic regression/IRT model (e.g., [1, 2, 21]) for each conditional probability table and, subsequently, manipulating the slope parameter. For underconfidence, we identified the highest probability p_L. We then transformed p_L into p'_L by making it smaller, according to the following formula:

$$p'_L = Lo^{-1}[Lo(p_L) - |\text{Normal}(0, \sigma)|] \ .$$

We made the smallest probability in the probability distribution Pr, p_S, higher by precisely the amount by which we decreased p_L, i.e.,

$$p'_S = p_S + p_L - p'_L \ .$$

We were by this guaranteed that the transformed parameters of the probability distribution Pr$'$ added up to 1.0.

Figure 2 shows the effect of introducing this biased noise. The middle two plots in the figure show overconfidence transformation and the bottom two show underconfidence. For overconfidence, in particular, the transformation is such that small probabilities are likely to become smaller and large probabilities are likely to become larger. Effectively, the distributions become more biased towards extreme probabilities.

We tested 30 versions of HEPAR II for each of the conditions (each network for a different standard deviation of the noise $\sigma \in < 0.0, 3.0 >$ with 0.1 increments) on all records of the HEPAR data set and computed HEPAR II's diagnostic accuracy. We plotted this accuracy in Figure 3 as a function of σ for different values of window size w. The left plot is for the overconfidence and the right plot is for the underconfidence condition.

It is clear that HEPAR II's diagnostic performance deteriorates with biased noise as well. The results are qualitatively similar to those in Figure 1, although performance under overconfidence bias degraded more slowly with the amount of noise than performance under symmetric noise, which, in turn degraded more slowly than performance under underconfidence.

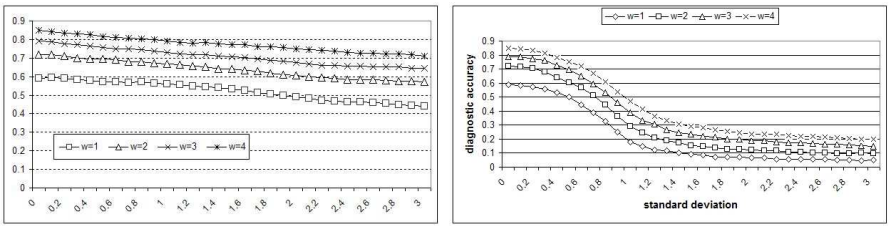


Fig. 3. The diagnostic accuracy of HEPAR II for various window sizes as a function of the amount of biased noise (expressed by σ). Overconfidence (left plot) and underconfidence (right plot) [10].

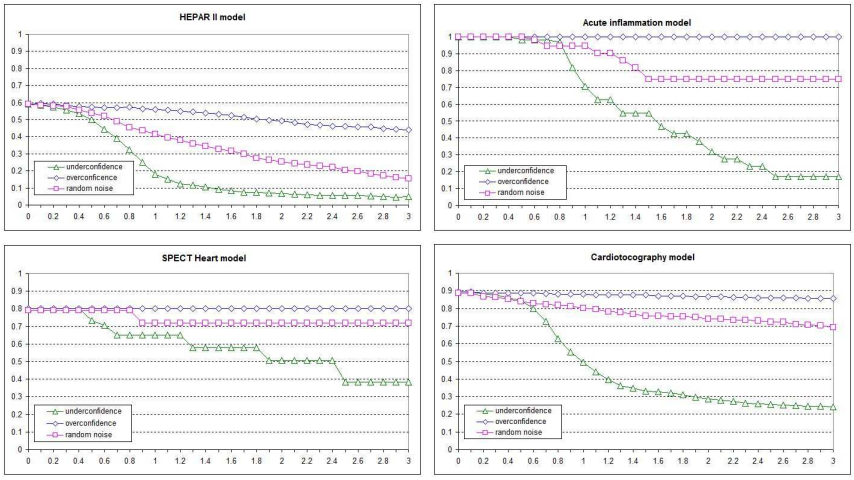


Fig. 4. The diagnostic accuracy of the four models (clock-wise HEPAR II, ACUTE INFLAMMATION, SPECT HEART and CARDIOTOGRAPHY) as a function of the amount of biased and unbiased noise, window $w = 1$.

We repeated this experiment for the three networks from the Irvine repository. Figure 4 shows the accuracy of HEPAR II and the three Irvine models as a function of the amount of biased and unbiased noise, window $w = 1$, on the same plot. The results are qualitatively identical: performance under underconfidence bias in all four cases degrades faster than performance under symmetric and overconfident noise.

It is interesting to note that here again for small values of σ, there is only a minimal effect of noise on performance.

4 Imprecision in parameters

Our next step was investigating how progressive rounding of a Bayesian network's probabilities affects its diagnostic performance. To that effect, we have

successively created various versions of models with different precision of parameters and tested the performance of these models.

For the purpose of our experiment, we used $n = 100$, 10, 5, 4, 3, 2, and 1, for the number of intervals in which the probabilities fall. And so, for $n = 10$, we divided the probability space into 10 intervals and each probability took one of 11 values, i.e., 0.0, 0.1, 0.2, ..., 0.9, and 1.0. For $n = 5$, each probability took one of six values, i.e., 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0. For $n = 2$, each probability took one of only three values, i.e., 0.0, 0.5, and 1.0. Finally, for $n = 1$, the smallest possible value of n, each probability was either 0.0 or 1.0. Figure 5 shows scatter plots of all 2,139 HEPAR II's parameters (horizontal axis) against their rounded values (vertical axis) for n equal to 10, 5, 2, and 1.

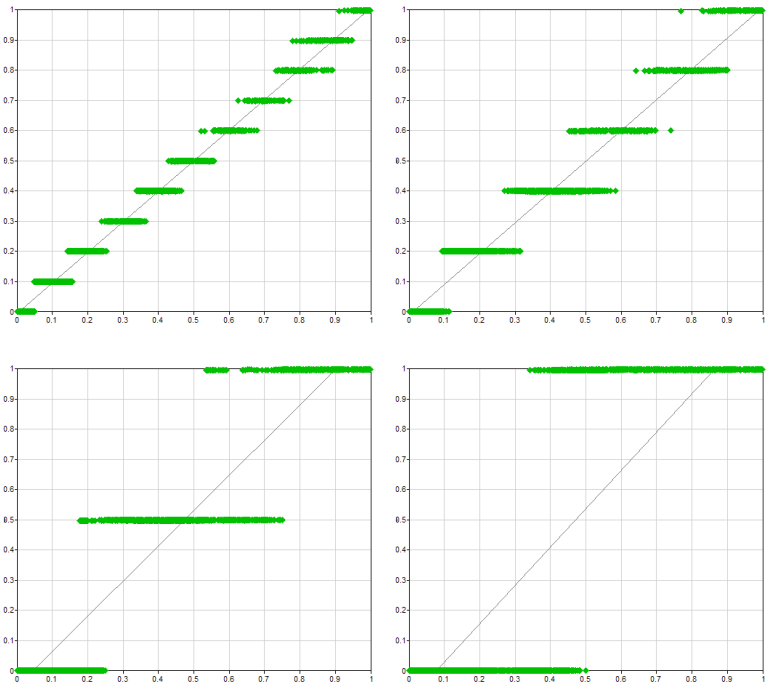
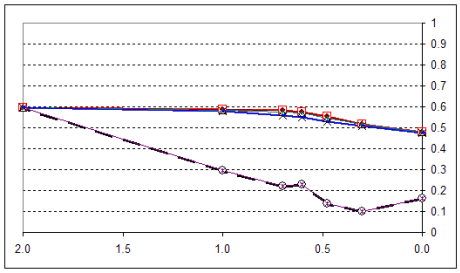


Fig. 5. Rounded vs. original probabilities for various levels of rounding accuracy.

Please note the drastic reduction in precision of the rounded probabilities, as pictured by the vertical axis. When $n = 1$, all rounded probabilities are either 0 or 1. Also, note that the horizontal bars in the scatter plot overlap. For example, in the upper-right plot ($n = 5$), we can see that an original probability $p = 0.5$ in HEPAR II got rounded sometimes to 0.4 and sometimes to 0.6. This is a simple consequence of the surrounding probabilities in the same distribution and the

necessity to make the sum of rounded probabilities add to 1.0, as guaranteed by the algorithm that we used for rounding probabilities.

We computed the diagnostic accuracy of various versions of HEPAR II, as produced by the rounding procedure. Figure 6 shows a summary of the results in both graphical and tabular format. The horizontal axis in the plot corresponds to the number of intervals n in logarithmic scale, i.e., value 2.0 corresponds to the rounding $n = 100$, and value 0 to the rounding $n = 1$. Intermediate points, for the other roundings can be identified in-between these extremes. The numerical accuracy reported in the table corresponds to the lower curve in the plot.



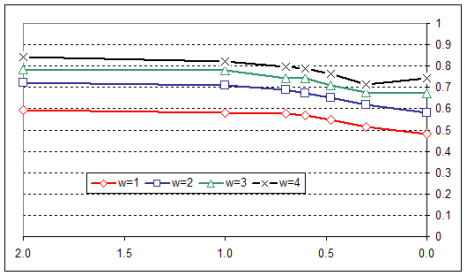
	accuracy	**# zeros**	**% zeros**
n=100	0.595	116	5%
n=10	0.295	400	19%
n=5	0.219	605	28%
n=4	0.230	754	35%
n=3	0.136	869	41%
n=2	0.100	1056	49%
n=1	0.161	1453	68%

Fig. 6. Diagnostic performance of HEPAR II as a function of logarithm of parameter accuracy and ε (w=1) [9].

It turns out that the strongly deteriorating accuracy is the effect of zeros in the probability distributions introduced by rounding. Please note that zero in probability theory is a special value. Once the probability of an event becomes zero, it can never change, no matter how strong the evidence for it. We addressed this problem by replacing all zeros introduced by the rounding algorithm by small ε probabilities and subtracting the introduced εs from the probabilities of the most likely outcomes in order to preserve the constraint that the sum should be equal to 1.0. While this caused a small distortion in the probability distributions (e.g., a value of 0.997 instead of 1.0 when $\varepsilon = 0.001$ and there were three induced zeros transformed into ε), it did not introduce sufficient difference to invalidate the precision loss. To give the reader an idea of what it entailed in practice, we will reveal the so far hidden information that the plots in Figure 5 were obtained for data with $\varepsilon = 0.001$.

The result of this modification was dramatic and is pictured by the upper curves in Figure 6, each line for a different value of ε. As can be seen, the actual value of ε did not matter too much (we tried three values: 0.0001, 0.001, and 0.01). In each case HEPAR II's performance was barely affected by rounding, even when there was just one interval, i.e., when all probabilities were either ε or $1 - \varepsilon$.

Our next experiment focused on the influence of precision in probabilities on HEPAR II's accuracy for windows of size 1, 2, 3, and 4. Figure 7 shows a summary of the results in both graphical and tabular format. The meaning of



	w=1	w=2	w=3	w=4
n=100	0.595	0.721	0.785	0.844
n=10	0.582	0.708	0.778	0.823
n=5	0.575	0.688	0.742	0.797
n=4	0.569	0.671	0.741	0.788
n=3	0.546	0.649	0.711	0.765
n=2	0.515	0.618	0.675	0.715
n=1	0.479	0.581	0.674	0.744

Fig. 7. Diagnostic performance of HEPAR II as a function of the logarithm of parameter accuracy and various window sizes [9].

the horizontal and vertical axes is the same as in Figure 6. We can see that the stability of HEPAR II's performance is similar for all window sizes.

We repeated the rounding experiment for the three networks from the Irvine repository. Figure 8 shows the accuracy of HEPAR II and the three Irvine models (window $w = 1$) as a function of the logarithm of parameter accuracy on the same plot. The results were qualitatively identical to those involving HEPAR II.

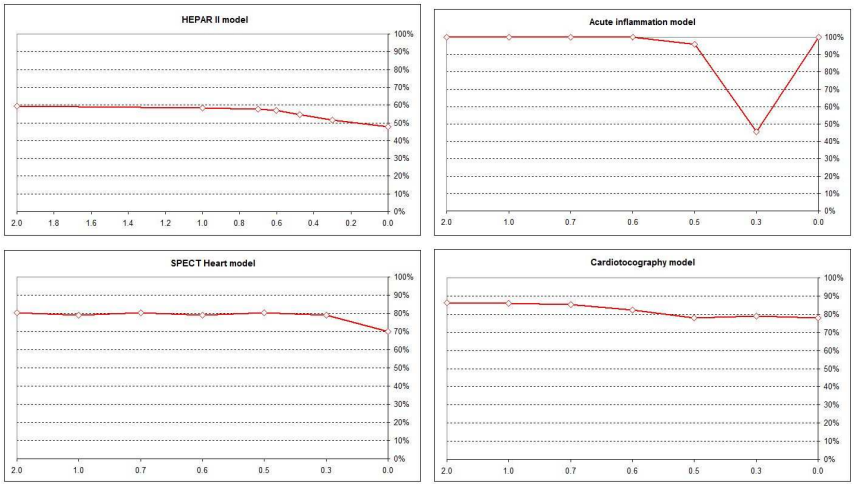


Fig. 8. The diagnostic accuracy of the four models (clock-wise HEPAR II, ACUTE INFLAMMATION, SPECT HEART and CARDIOTOGRAPHY) as a function of the logarithm of parameter accuracy, window $w = 1$.

5 Discussion

We described a series of experiments studying the influence of precision in parameters on model performance in the context of a practical medical diagnostic model, HEPAR II (these results was previously published in [9, 10, 17]), and three additional models based on real medical data from the Irvine Machine Learning Repository. We believe that the study was realistic in the sense of studying real models and focusing on a practical performance measure.

Our study has shown that the performance of all four models is sensitive to noise in numerical parameters, i.e., the diagnostic accuracy of the models decreases after introducing noise into their numerical parameters. For small to moderate amounts of noise, i.e., σ smaller than say 0.2, the effect of noise on accuracy was minimal. The effect of rounding the parameters was also minimal, giving some support to insensitivity of Bayesian network models to precision of their parameters.

We studied the influence of bias in parameters on model performance. Overconfidence bias had in our experiments a smaller negative effect on model performance than random noise. Underconfidence bias led to most serious deterioration of performance. While it is only a wild speculation that begs for further investigation, one might see our results as an explanation why humans tend to be overconfident rather than underconfident in their probability estimates. An interesting suggestion on the part of one of the reviewers was the link between bias, as we formulated it, and entropy. Models with parameters biased toward underconfidence have higher entropy and, thus, contain less information than models with symmetric noise or models biased toward overconfidence.

Our study of the influence of precision in parameters on model performance was inspired by the work of Clancey and Cooper [5], who conducted an experiment probing the sensitivity of MYCIN to the accuracy of its numerical specifications of degree of belief, certainty factors (CF). They applied a progressive roughening of CFs by mapping their original values onto a progressively coarser scale. The CF scale in MYCIN had 1,000 intervals ranging between 0 and 1,000. If this number was reduced to two, for example, every positive CF was replaced by the closest of the following three numbers: 0, 500, and 1,000. Roughening CFs to hundred, ten, five, three, and two intervals showed that MYCIN is fairly insensitive to their accuracy. Only when the number of intervals was reduced to three and two, there was a noticeable effect on the system performance.

Our results are somewhat different. It appears that the diagnostic accuracy of Bayesian network models is sensitive to imprecision in probabilities, if these are rounded. However, the main source of this sensitivity appears to be in rounding small probabilities to zero. When zeros introduced by rounding are replaced by very small non-zero values, imprecision resulting from rounding has minimal impact on Bayesian network model's performance.

Acknowledgments

Agnieszka Onisko was supported by the Białystok University of Technology grants W/WI/1/02 and S/WI/2/2008, by the MNiI (Ministerstwo Nauki i Informatyzacji) grant 3T10C03529, and by the Polish Committee for Scientific Research grant 4T11E05522. Marek Druzdzel was supported by the National Institute of Health under grant number U01HL101066-01, by the Air Force Office of Scientific Research grants F49620–00–1–0112, F49620–03–1–0187, and FA9550-06-1-0243, and by Intel Research.

While we are solely responsible for any remaining shortcomings of this paper, our work has benefitted from helpful comments and suggestions from several individuals, of whom we would like to thank in particular Greg Cooper and Linda van der Gaag. Anonymous reviewers asked several excellent questions and offered suggestions that led to improvements of the paper.

All Bayesian network models in this paper were created and tested using SMILE, an inference engine, and GeNIe, a development environment for reasoning in graphical probabilistic models, both developed at the Decision Systems Laboratory and available at `http://genie.sis.pitt.edu/`.

References

1. Almond, R.G., Dibello, L.V., Jenkins, F., Mislevy, R., Senturk, D., Steinberg, L., Yan, D.: Models for conditional probability tables in educational assessment. In: Jaakkola, T., Richardson, T. (eds.) Artificial Intelligence and Statistics. pp. 137–143. Morgan Kaufmann (2001)
2. Almond, R.G., Dibello, L.V., Moulder, B., Zapata-Rivera, J.D.: Modeling diagnostic assessment with Bayesian networks. Journal of Educational Measurement 44(4), 341–359 (2007)
3. Chan, H., Darwiche, A.: When do numbers really matter? Journal of Artificial Intelligence Research 17, 265–287 (2002)
4. Cios, K.J., Kurgan, L.A.: UCI machine learning repository (2011), http://archive.ics.uci.edu/ml
5. Clancey, W.J., Cooper, G.: Uncertainty and evidential support. In: Buchanan, B.G., Shortliffe, E.H. (eds.) Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, chap. 10, pp. 209–232. Addison-Wesley, Reading, MA (1984)
6. Coupé, V.H.M., van der Gaag, L.: Practicable sensitivity analysis of Bayesian belief networks. In: Prague Stochastics '98 — Proceedings of the Joint Session of the 6th Prague Symposium of Asymptotic Statistics and the 13th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes. pp. 81–86. Union of Czech Mathematicians and Physicists (1998)
7. Coupé, V.H.M., van der Gaag, L.C.: Properties of sensitivity analysis of Bayesian belief networks. Annals of Mathematics and Artificial Intelligence 36, 323–356 (2002)
8. Czerniak, J., Zarzycki, H.: Application of rough sets in the presumptive diagnosis of urinary system diseases. In: Artifical Inteligence and Security in Computing Systems, ACS'2002 9th International Conference. pp. 41–51. Kluwer Academic Publishers (2003)

9. Druzdzel, M.J., Oniśko, A.: Are Bayesian networks sensitive to precision of their parameters? In: S.T. Wierzchoń, M.K., Michalewicz, M. (eds.) Proceedings of the Intelligent Information Systems Conference (XVI). pp. 35–44. Academic Publishing House EXIT, Warsaw, Poland (2008)
10. Druzdzel, M.J., Oniśko, A.: The impact of overconfidence bias on practical accuracy of Bayesian network models: An empirical study. In: Working Notes of the 2008 Bayesian Modelling Applications Workshop, Special Theme: How Biased Are Our Numbers? Part of the Annual Conference on Uncertainty in Artificial Intelligence (UAI-2008). Helsinki, Finland (2008)
11. van der Gaag, L.C., Renooij, S.: Analysing sensitivity data from probabilistic networks. In: Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2001). pp. 530–537. Morgan Kaufmann Publishers, San Francisco, CA (2001)
12. Henrion, M., Breese, J.S., Horvitz, E.J.: Decision Analysis and Expert Systems. AI Magazine 12(4), 64–91 (Winter 1991)
13. Kahneman, D., Slovic, P., Tversky, A. (eds.): Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge (1982)
14. Kjaerulff, U., van der Gaag, L.C.: Making sensitivity analysis computationally efficient. In: Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2000). pp. 317–325. Morgan Kaufmann Publishers, San Francisco, CA (2000)
15. Middleton, B., Shwe, M., Heckerman, D., Henrion, M., Horvitz, E., Lehmann, H., Cooper, G.: Probabilistic diagnosis using a reformulation of the INTERNIST–1/QMR knowledge base: II. Evaluation of diagnostic performance. Methods of Information in Medicine 30(4), 256–267 (1991)
16. Morgan, M.G., Henrion, M.: Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press, Cambridge (1990)
17. Oniśko, A., Druzdzel, M.J.: Effect of imprecision in probabilities on Bayesian network models: An empirical study. In: Working notes of the European Conference on Artificial Intelligence in Medicine (AIME-03): Qualitative and Model-based Reasoning in Biomedicine. Protaras, Cyprus (October 18–22 2003)
18. Oniśko, A., Druzdzel, M.J., Wasyluk, H.: Learning Bayesian network parameters from small data sets: Application of Noisy-OR gates. International Journal of Approximate Reasoning 27(2), 165–182 (2001)
19. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers, Inc., San Mateo, CA (1988)
20. Pradhan, M., Henrion, M., Provan, G., del Favero, B., Huang, K.: The sensitivity of belief networks to imprecise probabilities: An experimental investigation. Artificial Intelligence 85(1–2), 363–397 (Aug 1996)
21. Rijmen, F.: Bayesian networks with a logistic regression model for the conditional probabilities. International Journal of Approximate Reasoning In press
22. Marques de Sá, J., Bernardes, J., Ayres de Campos, D.: UCI Machine Learning Repository (2011), http://archive.ics.uci.edu/ml
23. Shwe, M., Middleton, B., Heckerman, D., Henrion, M., Horvitz, E., Lehmann, H., Cooper, G.: Probabilistic diagnosis using a reformulation of the INTERNIST–1/QMR knowledge base: I. The probabilistic model and inference algorithms. Methods of Information in Medicine 30(4), 241–255 (1991)